Weakly supervised discriminative localization and classification: a joint learning process

Minh Hoai Nguyen[1] Lorenzo Torresani[2] Fernando de la Torre[1]
Carsten Rother[3]

Technical Report – CMU-RI-TR-09-29
The Robotics Institute, Carnegie Mellon University

July 15, 2009

[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]Dartmouth College, Hanover, NH, USA
[3]Microsoft Research Cambridge, Cambridge, UK

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **15 JUL 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **Weakly supervised discriminative localization and classification: a joint learning process**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**24**	

Abstract

Visual categorization problems, such as object classification or action recognition, are increasingly often approached using a detection strategy: a classifier function is first applied to candidate subwindows of the image or the video, and then the maximum classifier score is used for class decision. Traditionally, the subwindow classifiers are trained on a large collection of examples manually annotated with masks or bounding boxes. The reliance on time-consuming human labeling effectively limits the application of these methods to problems involving very few categories. Furthermore, the human selection of the masks introduces arbitrary biases (e.g. in terms of window size and location) which may be suboptimal for classification.

In this report we propose a novel method for learning a discriminative subwindow classifier from examples annotated with binary labels indicating the presence of an object or action of interest, but *not* its location. During training, our approach simultaneously localizes the instances of the positive class and learns a subwindow SVM to recognize them. We extend our method to classification of time series by presenting an algorithm that localizes the most discriminative set of temporal segments in the signal. We evaluate our approach on several datasets for object and action recognition and show that it achieves results similar and in many cases superior to those obtained with full supervision.

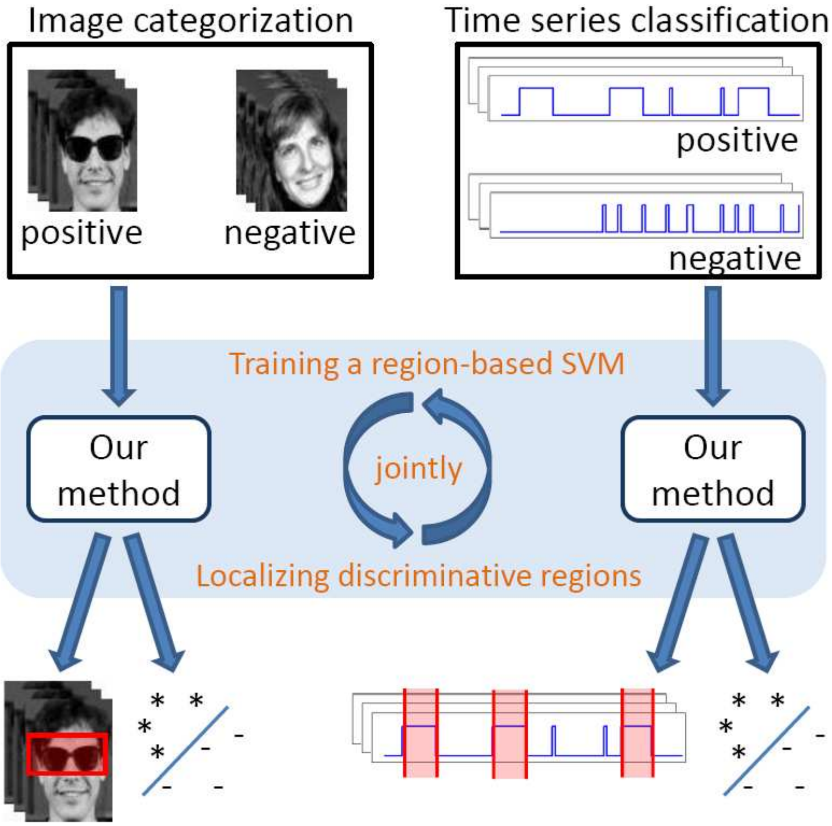
Image categorization
positive
negative
Time series classification
positive
negative
Training a region-based SVM
Our method
jointly
Our method
Localizing discriminative regions

Figure 1: A unified framework for image categorization and time series classification from weakly labeled data. Our method simultaneously localizes the regions of interest in the examples and learns a region-based classifier, thus building robustness to background and uninformative signal.

1 Introduction

Object categorization systems aim at recognizing the classes of the objects present in an image, independently of the background. Early computer vision methods for object categorization attempted to build robustness to background clutter by using image segmentation as preprocessing. It was hoped that segmentation methods could partition images into their high-level constituent parts, and categorization could then be simply carried out as recognition of the object classes corresponding to the segments. This naive strategy to categorization floundered on the challenges presented by bottom-up image segmentation. The difficulty of partitioning an image into objects purely based on low-level cues is now well understood and it has led in recent years to a flourishing of methods where bottom-up segmentation is assisted by concurrent top-down recognition [31, 17, 4, 27]. However, the application of these methods has been limited in practice by *a)* the challenges posed by the acquisition of detailed ground truth segmentations needed to train these systems, and *b)* the high computational complexity of semantic segmentation, which requires solving the classification problem at the pixel-level. An efficient alternative is provided by object detection methods, which can perform object localization without requiring pixel-level segmentation. Object detection algorithms operate by evaluating a classifier function at many different subwindows of the image and then predicting the object presence in subwindows with high-score. This methodology has been applied with great success to a wide variety of object classes [29, 8, 7]. Recent work [15] has shown that efficient computation of classification maxima over all possible subwindows of an image is even possible for highly sophisticated classifiers, such as SVMs with spatial pyramid kernels. Although great advances have been made in terms of reducing the computational complexity of object detection algorithms, their accuracy has remained dependent on the amount of human-annotated data available to train them. Subwindows (or bounding boxes) are obviously less-time consuming to collect than detailed segmentations. However, the dependence on human work for training inevitably limits the scalability of these methods. Furthermore, not only the amount of ground truth data but also the characteristics of the human selections may affect the detection. For example, it has been shown [8] that the specific size and location of the selections may have a significant impact on performance. In some cases, including a margin around the bounding box of the training selections will lead to better detection because of statistical correlation between the appearance of the region surrounding the object (often referred to as the "spatial context") and the category of the object (e.g. cars tend to appear on roads). However, it is rather difficult to tune the amount of context to include for optimal classification. The problem is even more acute for the case of categorization of time series data. Consider the task of automatically monitoring the behavior of an animal based on its body movement. It is safe to believe that the intrinsic differences between the distinct animal activities (e.g. drinking, exploring, etc.) do not appear continuously in the examples but are rather associated to specific movement patterns (e.g. the turning of the head, a short fast-pace walk, etc.) possibly occurring multiple times in the sequences. Thus, as for the case of object categorization, classification based on comparisons of the whole signals is unlikely to yield good performance. However, if we asked a person to localize the most discriminative patterns in such sequences, we would obtain highly subjective annotations, unlikely to be optimal for the training of a classifier.

In this report we propose a novel learning framework that simultaneously localizes the most discriminative subwindows in the data and learns a classifier to distinguish them. Our algorithm requires only the class labels as annotation for the training examples, and thus eliminates the high cost and arbitrariness of human ground truth selections. In the case of object categorization, our method optimizes an SVM classification objective with respect to both the classifier parameters and the subwindows containing the object of interest in the positive image examples. In the case of classification of time series, we relax the subwindow contiguity constraint in order to discover discriminative patterns which may occur discontinuously over the observation period. Specifically, we allow the discriminative patterns to occur in at most k disjoint time-intervals, where k is a problem-dependent tunable parameter of our system. The algorithm solves for the locations and durations of these intervals while learning the SVM classifier. We demonstrate our approach on several object and activity recognition datasets and show that our weakly-supervised classifiers consistently match and often surpass the accuracy of SVMs trained under full supervision.

2 Related Work

Most prior work on weakly supervised object localization and classification is based on the use of region or part-based generative models. Fergus et al. [12] represent objects as flexible constellation of parts by learning probabilistic models of both the appearance as well as the mutual position of the parts. Parts are selected from points found by a feature detector. Classification of a test image is performed in a Bayesian fashion by evaluating the detected features using the learned model. The performance of this system rests completely on the ability of the feature detector to fire consistently at points corresponding to the learned parts of the model. Russell et al. [23] instead propose a fully-unsupervised algorithm to discover objects and associated segments from a large collection of images. Multiple segmentations are computed from each image by varying the parameters of a segmentation method. The key-assumption is that each object instance is correctly segmented at least once and that the features of correct segments form object-specific coherent clusters discoverable using latent topic models from text analysis. Although the algorithm is shown to be able to discover many different types of objects, its effectiveness as a categorization technique is unclear. Cao and Fei-Fei [5] further extend the latent topic model by assuming that a single topic model is responsible for generating the image patches within each region of the image, thus enforcing spatial coherence within each segment. Todorovic and Ahuja [26] describe a system that learns tree-based representations of multiscale image segmentations via a subtree matching algorithm. A multitude of algorithms based on Multiple Instance Learning (MIL) have been recently proposed for training object classifiers with weakly supervised data (see [19, 30, 2, 6] for a sampling of these techniques). Most of these methods view images as bags of segments, traditionally computed using bottom-up segmentation or fixed partitioning of the image into blocks. Then MIL trains a discriminative binary classifier predicting the class of segments, under the assumption that each positive training image contains at least one true-positive segment (corresponding to the object of interest), while negative training images contain none. However, these approaches incur in the same problem faced by the early segmentation-based recognition systems: segmentation from low-level cues is often unable to provide semantically correct segments. Galleguillos et al. [13] attempt to circumvent this problem by providing multiple segmentations to the MIL learning algorithm in the hope one of them is correct. The approach we propose does not rely on unreliable segmentation methods as preprocessing. Instead, it performs localization while training the classifier. Our work can also be viewed as an extension of feature selection methods, in which different features are selected for each example. The idea of *joint* feature selection and classifier optimization has been proposed before, but always in combination with strongly labeled data. Schweitzer [24] proposes a linear time algorithm to select jointly a subset of pixels and a set of eigenvectors that minimize the Rayleigh quotient in Linear Discriminant Analysis. Nguyen et al. [20] propose a convex formulation to simultaneously select the most discriminative pixels and optimize the SVM parameters. However, both aforementioned methods require the training data to be well aligned and the same set of pixels is selected for every image. Felzenszwalb et al. [11] describe Latent SVM, a powerful classification framework based on a deformable part model. However, also this method requires knowing the bounding boxes of foreground objects during training. Finally, Blaschko and Lampert [3] use *supervised* structured learning to improve the localization accuracy of SVMs.

The literature on weakly supervised or unsupervised localization and categorization applied to time series is fairly limited compared to the object recognition case. Zhong et al. [32] detect unusual activities in videos by clustering equal-length segments extracted from the video. The segments falling in isolated clusters are classified as abnormal activities. Fanti et al. [10] describe a system for unsupervised human motion recognition from videos. Appearance and motion cues derived from feature tracking are used to learn graphical models of actions based on triangulated graphs. Niebles et al. [21] tackle the same problem but represent each video as a bag of video words, i.e. quantized descriptors computed at spatial-temporal interest points. An EM algorithm for topic models is then applied to discover the latent topics corresponding to the distinct actions in the dataset. Localization is obtained by computing the MAP topic of each word.

3 Localization–classification SVM

In this section we first propose an algorithm to simultaneously localize objects of interest and train an SVM. We then extend it to classification of time series data by presenting an efficient algorithm to identify in the signal an optimal set of discriminative segments, which are not constrained to be contiguous.

3.1 The learning objective

Assume we are given a set of positive training images $\{\mathbf{d}_i^+\}$ and a set of negative training images $\{\mathbf{d}_i^-\}$ corresponding to weakly labeled data with labels indicating for each example the presence or absence of an object of interest. Let $\mathcal{LS}(\mathbf{d})$ denote the set of all possible subwindows of image $\mathbf{d}$. Given a subwindow $\mathbf{x} \in \mathcal{LS}(\mathbf{d})$, let $\varphi(\mathbf{x})$ be the feature vector computed from the image subwindow. We learn an SVM for joint localization and classification by solving the following constrained optimization:

$$\underset{\mathbf{w},b}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||^2, \tag{1}$$

$$\text{s.t.} \max_{\mathbf{x}\in\mathcal{LS}(\mathbf{d}_i^+)} \{\mathbf{w}^T\varphi(\mathbf{x}) + b\} \geq 1 \ \forall i, \tag{2}$$

$$\max_{\mathbf{x}\in\mathcal{LS}(\mathbf{d}_i^-)} \{\mathbf{w}^T\varphi(\mathbf{x}) + b\} \leq -1 \ \forall i. \tag{3}$$

The constraints appearing in this objective state that each positive image must contain at least one subwindow classified as positive, and that *all* subwindows in each negative image must be classified as negative. The goal is then to maximize the margin subject to these constraints. By optimizing this problem we obtain an SVM, i.e. parameters $(\mathbf{w}, b)$, that can be used for localization and classification.

Given a new testing image $\mathbf{d}$, localization and classification are done as follows. First, we find the subwindow $\hat{\mathbf{x}}$ yielding the maximum SVM score:

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathcal{LS}(\mathbf{d})} \mathbf{w}^T \varphi(\mathbf{x}). \tag{4}$$

If the value of $\mathbf{w}^T \varphi(\hat{\mathbf{x}}) + b$ is positive, we report $\hat{\mathbf{x}}$ as the detected object for the test image. Otherwise, we report no detection.

As in the traditional formulation of SVM, the constraints are allowed to be violated by introducing slack variables:

$$\underset{\mathbf{w},b}{\text{minimize}} \ \frac{1}{2}||\mathbf{w}||^2 + C \sum_i \alpha_i + C \sum_i \beta_i, \tag{5}$$

$$\text{s.t.} \ \max_{\mathbf{x} \in \mathcal{LS}(\mathbf{d}_i^+)} \{\mathbf{w}^T \varphi(\mathbf{x}) + b\} \geq 1 - \alpha_i \ \forall i, \tag{6}$$

$$\max_{\mathbf{x} \in \mathcal{LS}(\mathbf{d}_i^-)} \{\mathbf{w}^T \varphi(\mathbf{x}) + b\} \leq -1 + \beta_i \ \forall i, \tag{7}$$

$$\alpha_i \geq 0, \ \beta_i \geq 0 \ \forall i.$$

Here, C is the parameter controlling the trade-off between having a large margin and less constraint violation.

3.2 Optimization

Our objective is in general non-convex. We propose optimization via a coordinate descent approach that alternates between optimizing the objective w.r.t. parameters $(\mathbf{w}, b, \{\alpha_i\}, \{\beta_i\})$ and finding the subwindows of images $\{\mathbf{d}_i^+\} \cup \{\mathbf{d}_i^-\}$ that maximize the SVM scores. However, since the cardinality of the sets of all possible subwindows may be very large, special treatment is required for constraints of type (7). We use *constraint generation* to handle these constraints: $\mathcal{LS}(\mathbf{d}_i^-)$ is iteratively updated by adding the most violated constraint at every step. Although constraint generation has exponential running time in the worst case, it often works well in practice.

The above optimization requires at each iteration to localize the subwindow maximizing the SVM score in each image. Thus, we need a very fast localization procedure. For this purpose, we adopt the representation and algorithm described in [15]. Images are represented as bags of visual words obtained by quantizing SIFT descriptors [18] computed at random locations and scales. For quantization, we use a visual dictionary built by applying K-means clustering to a set of descriptors extracted from the training images [25]. The set of possible subwindows for an image is taken to be the set of axis-aligned rectangles. The feature vector $\varphi(\mathbf{x})$ is the histogram of visual words associated with descriptors inside rectangle $\mathbf{x}$. Lampert et al. [15] showed that, when using this image representation, the search for the rectangle maximizing the SVM score can be executed efficiently by means of a branch-and-bound algorithm.

3.3 Extension to time series

As in the case of image categorization, even for time series the global statistics computed from the entire signal may yield suboptimal classification. For example, the differences between two classes of temporal signals may not be visible over the entire observation period. However, unlike in the case of images where objects often appear as fully-connected regions, the patterns of interest in temporal signals may not be contiguous. This raises a technical challenge when extending the learning formulation of Eq. (5) to time series classification: how to efficiently search for sets of non-contiguous discriminative segments? In this section we describe a representation of temporal signals and a novel efficient algorithm to address this challenge.

3.3.1 Representation of time series

Time series can be represented by descriptors computed at spatial-temporal interest points [16, 9, 21]. As in the case of images, sample descriptors from training data can be clustered to create a visual-temporal vocabulary [9]. Subsequently, each descriptor is represented by the ID of the corresponding vocabulary entry and the frame number at which the point is detected. In this work, we define a k-*segmentation* of a time series as a set of k disjoint time-intervals, where k is a tunable parameter of the algorithm. Note that it is possible for some intervals of a k-segmentation to be empty. Given a k-segmentation $\mathbf{x}$, let $\varphi(\mathbf{x})$ denote the histogram of visual-temporal words associated with interest points in $\mathbf{x}$. Let C_i denote the set of words occurring at frame i. Let $a_i = \sum_{c \in C_i} w_c$ if C_i is non-empty, and $a_i = 0$ otherwise. a_i is the weighted sum of words occurring in frame i where word c is weighted by SVM weight w_c. From these definitions it follows that $\mathbf{w}^T\varphi(\mathbf{x}) = \sum_{i \in \mathbf{x}} a_i$. For fast localization of discriminative patterns in time series we need an algorithm to efficiently find the k-segmentation maximizing the SVM score $\mathbf{w}^T\varphi(\mathbf{x})$. Indeed, this optimization can be solved globally in a very efficient way. The following section describes the algorithm. In the appendix, we prove the optimality of the solution produced by this algorithm.

3.3.2 An efficient localization algorithm

Let n be the length of the time signal and $\mathcal{I} = \{[l, u] : 1 \leq l \leq u \leq n\}$ be the set of all subintervals of $[1, n]$. For a subset $S \subseteq \{1, \cdots, n\}$, let $f(S) = \sum_{i \in S} a_i$. Maximization of $\mathbf{w}^T \varphi(\mathbf{x})$ is equivalent to:

$$\underset{I_1, \ldots, I_k}{\text{maximize}} \sum_{j=1}^{k} f(I_j) \text{ s.t. } I_i \in \mathcal{I} \;\&\; I_i \cap I_j = \phi \;\forall i \neq j. \tag{8}$$

This problem can be optimized very efficiently using Algo. 1 presented below.

Algorithm 1 Find best k disjoint intervals that optimize (8)

Input: $a_1, \cdots, a_n$, $k \geq 1$.

Output: a set $\mathcal{X}^k$ of best k disjoint intervals.

```
1: $\mathcal{X}^0 := \phi$.
2: for $m = 0$ to $k - 1$ do
3:     $J_1 := \arg\max_{J \in \mathcal{I}} f(J)$ s.t. $J \cap S = \phi \; \forall S \in \mathcal{X}^m$.
4:     $J_2 := \arg\max_{J \in \mathcal{I}} -f(J)$ s.t. $J \subset S \in \mathcal{X}^m$.
5:     if $f(J_1) \geq -f(J_2)$ then
6:         $\mathcal{X}^{m+1} := \mathcal{X}^m \cup \{J_1\}$
7:     else
8:         Let $S \in \mathcal{X}^m : J_2 \subset S$. $S$ is divided into three disjoint intervals: $S = S^- \cup J_2 \cup S^+$.
9:         $\mathcal{X}^{m+1} := (\mathcal{X}^m - \{S\}) \cup \{S^-, S^+\}$
```

This algorithm progressively finds the set of m intervals (possibly empty) that maximize (8) for $m = 1, \cdots, k$. Given the optimal set of m intervals, the optimal set of $m + 1$ intervals is obtained as follows. First, find the interval J_1 that has maximum score $f(J_1)$ among the intervals that do not overlap with any currently selected interval (line 3). Second, locate J_2, the worst subinterval of all currently selected intervals, i.e. the subinterval with lowest score $f(J_2)$ (line 4). Finally, the optimal set of $m + 1$ intervals is constructed by executing either of the following two operations, depending on which one leads to the higher objective:

1. Add J_1 to the optimal set of m intervals (line 6);
2. Break the interval of which J_2 is a subinterval into three intervals and remove J_2 (line 9).

Algo. 1 assumes J_1 and J_2 can be found efficiently. This is indeed the case. We now describe the procedure for finding J_1. The procedure for finding J_2 is similar.

Let $\overline{\mathcal{X}^m}$ denote the relative complement of $\mathcal{X}^m$ in $[1, n]$, i.e. $\overline{\mathcal{X}^m}$ is the set of intervals such that the "union" of the intervals in $\mathcal{X}^m$ and $\overline{\mathcal{X}^m}$ is the interval $[1, n]$. Since $\mathcal{X}^m$ has at most m elements, $\overline{\mathcal{X}^m}$ has at most $m + 1$ elements. Since J_1 does not intersect with any interval in $\mathcal{X}^m$, it must be a subinterval of an interval of $\overline{\mathcal{X}^m}$. Thus, we can find J_1 as $J_1 = \arg\max_{S \in \overline{\mathcal{X}^m}} f(J^S)$ where:

$$J^S = \arg\max_{J \subseteq S} f(J). \tag{9}$$

Eq. (9) is a basic operation that is needed to be performed repeatedly: finding a subinterval of an interval that maximizes the sum of elements in that subinterval. This operation can be performed by Algo. 2 below with running time complexity $\mathbf{O}(n)$. Note that the result of executing (9) can be cached; we do not need to recompute J^S for many S at each iteration of Algo. 1. Thus the total running complexity of Algo. 1 is $\mathbf{O}(nk)$. Algo. 1 guarantees to produce a globally optimal solution for (8) (see the appendix).

Algorithm 2 Find the best subinterval

```
Input: $a_1, \cdots, a_n$, an interval $[l, u] \subset [1, n]$.
Output: $[sl, su] \subset [l, u]$ with maximum sum of elements.
$b_0 := 0$.
for $m = 1$ to $n$ do
    $b_m := b_{m-1} + a_m$. //compute integral image
$[sl, su] := [0, 0]; val := 0$. //empty subinterval
$\widehat{m} := l - 1$. //index for minimum element so far
for $m = l$ to $u$ do
    if $b_m - b_{\widehat{m}} > val$ then
        $[sl, su] := [\widehat{m} + 1; m]; val := b_m - b_{\widehat{m}}$
    else if $b_m < b_{\widehat{m}}$ then
        $\widehat{m} := m$. //keep track of the minimum element
```

4 Experiments

This section describes experiments on several datasets for object categorization and time series classification.

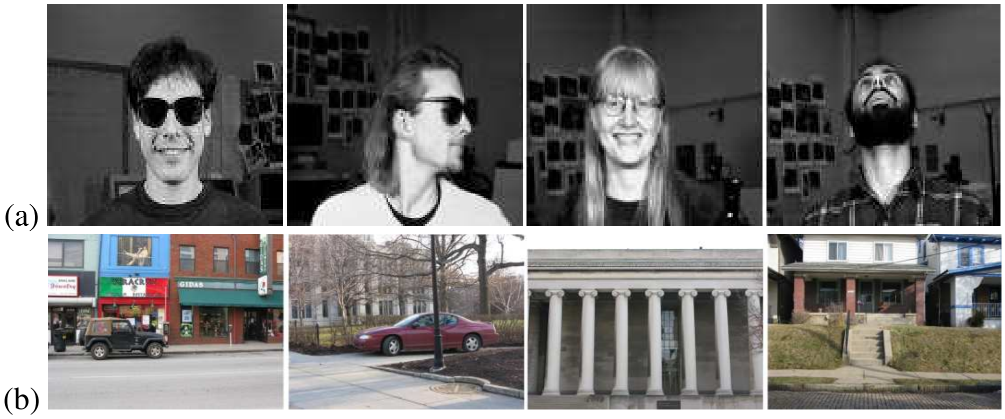

Figure 2: Examples taken from (a) the CMU Face Images and (b) the street scene dataset.

4.1 Object localization and categorization

4.1.1 Experiments on car and face datasets

This subsection presents evaluations on two image collections. The first experiment was performed on CMU Face Images, a publicly available dataset from the UCI machine learning repository[1]. This database contains 624 face images of 20 people with different expressions and poses. The subjects wear sunglasses in roughly half of the images. Our classification task is to distinguish between the faces with sunglasses and the faces without sunglasses. Some image examples from the database are given in Fig. 2(a). We divided this image collection into disjoint training and testing subsets. Images of the first 8 people are used for training while images of the last 12 people are reserved for testing. Altogether, we had 254 training images (126 with glasses and 128 without glasses) and 370 testing images (185 examples for both the positive and the negative class).

The second experiment was performed on a dataset collected by us. Our collection contains 400 images of street scenes. Half of the images contain cars and half of them do not. This is a challenging dataset because the appearance of the cars in the images varies in shape, size, grayscale intensity, and location. Furthermore, the cars occupy only a small portion of the images and may be partially occluded by other objects. Some examples of images from this dataset are shown in Fig. 2(b). Given the limited amount of examples available, we applied 4-fold

[1] http://archive.ics.uci.edu/ml/datasets/CMU+Face+Images

cross validation to obtain an estimate of the performance.

Each image is represented by a set of 10,000 local SIFT descriptors [18] selected at random locations and scales. The descriptors are quantized using a dictionary of 1,000 visual words obtained by applying hierarchical K-means [22] to 100,000 training descriptors.

In order to speed up the learning, an upper constraint on the rectangle size is imposed. In the first experiment, as the image size is 120×128 and the sizes of sunglasses are relative small, we restrict the height and width of permissible rectangles to not exceed 30 and 50 pixels, respectively. Similarly, for the second experiment, we constrain permissible rectangles to have height and width no larger than 300 and 500 pixels, respectively (c.f. image size of 600×800).

We compared our approach to several competing methods. *SVM* denotes a traditional SVM approach in which each image is represented by the histogram of the words in the whole image. *BoW* is the *bag-of-words* method [22] in the implementation of [28]. It uses a 10-nearest neighbor classifier. We also benchmark our method against *SVM-FS* [15], a fully supervised method requiring ground truth subwindows during training (FS stands for fully supervised). *SVM-FS* trains an SVM using ground truth bounding boxes as positive examples and ten random rectangles from each negative image for negative data.

Tab. 1 shows the classification performance measured using both the accuracy rates and the areas under the ROCs. Note that our approach outperforms not only *SVM* and *BoW* (which are based on global statistics), but also *SVM-FS*, which is a fully supervised method requiring the bounding boxes of the objects during training. This suggests that the boxes tightly enclosing the objects of interest are not always the most discriminative regions. Our method automatically localizes the subwindows that are most discriminative for classification. Fig. 3(a) shows discriminative detection on a few face testing examples. Sunglasses are the distinguishing elements between positive and negative classes. Our algorithm successfully discovers such regions and exploits them to improve the classification performance. Fig. 3(b) shows some examples of car localization. Parts of the road below the cars tend to be included in the detection output. This suggests that the appearance of roads is a contextual indication for the presence of cars. Fig. 4 displays several difficult cases where our method does not provide good localization of the objects.

SVM, *SVM-FS*, and our proposed method require tuning of a single parameter, C, controlling the trade-off between a large margin and less constraint violation. This parameter is tuned using 4-fold cross validation on training data. The parameter sweeping is done exactly in the same fashion for all algorithms.

Dataset	Measure	BoW	SVM	SVM-FS	Ours
Faces	Acc. (%)	80.11	82.97	86.79	**90.0**
	ROC Area	n/a	0.90	0.94	**0.96**
Cars	Acc. (%)	77.5	80.75	81.44	**84.0**
	ROC Area	n/a	0.86	0.88	**0.90**

Table 1: Comparison results on the CMU Face and car datasets. BoW: bag of words approach [22]. SVM: SVM using global statistics. SVM-FS [15] requires bounding boxes of foreground objects during training. Our method is significantly better than the others, and it outperforms even the algorithm using strongly labeled data.

Optimizing (5) is an iterative procedure, where each iteration involves solving a convex quadratic programming problem. Our implementation uses `CVX`, a package for specifying and solving convex programs [14]. We found that our algorithm generally converges within 100 iterations of coordinate descent.

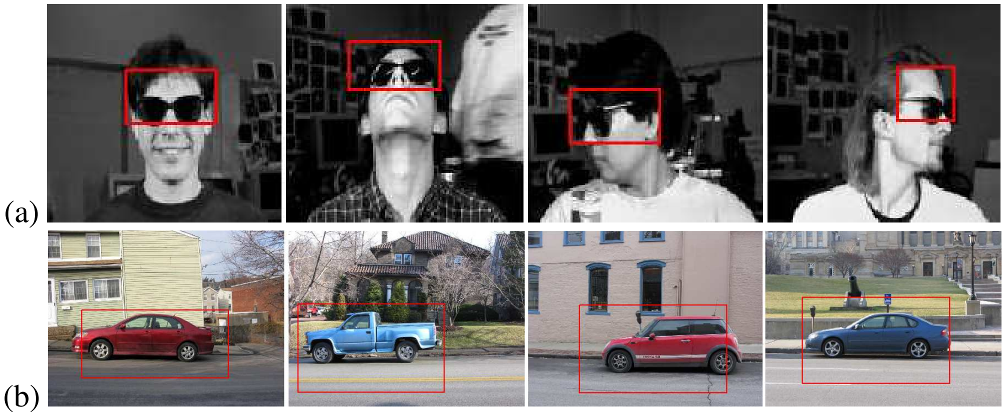

Figure 3: Localization of (a) sunglasses and (b) cars on test images. Note how the road below the cars is partially included in the detection output. This indicates that the appearance of road serves as a contextual indication for the presence of cars.

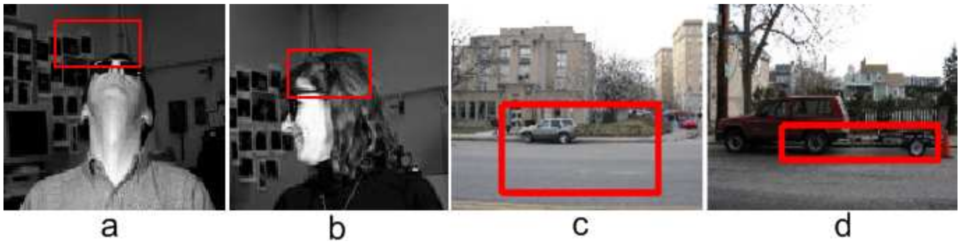

Figure 4: Difficult cases for localization. a, b: sunglasses are not clearly visible in the images. c: the foreground object is very small. d: misdetection due to the presence of the trailer wheel.

4.1.2 Experiments on Caltech-4

This subsection describes an experiment on the publicly available[2] Caltech-4 dataset. This collection contains images of different categories: airplanes_side, background, cars_bard, faces, and motorbikes_side. We consider binary classification tasks where the goal is to distinguish one of the four object classes (airplanes_side, cars_bard, faces, and motorbikes_side) from the background clutter class. In this experiment, we randomly sample a set of 100 images from each class for training. The set of the remaining images is split into equal-size testing and validation sets. The validation data is used for parameter tuning.

Tab. 2 shows the results of this experiment. As shown, *SVM-FS*, a method that requires bounding boxes of the foreground objects for training, does not perform as well as *SVM* which is based on global statistics from the whole image. This result suggests that contextual information is very important for classification tasks on this dataset. Indeed, it is easy to verify by visual inspection that the image backgrounds here often provide very strong categorization cues (see e.g. the almost constant background of the face images). As a result our method cannot provide any significant advantage on this dataset. However note that, unlike SVM-FS, our joint localization and classification does not harm the classification performance as our algorithm automatically learns the importance of contextual information and uses large subwindows for recognition.

4.2 Classification of time series data

This section describes our classification experiments on time series datasets.

[2] http://www.robots.ox.ac.uk/∼vgg/data3.html

Class	Measure	BoW	SVM	SVM-FS	Ours
Airplanes	Acc. (%)	89.74	**96.05**	89.40	**96.05**
	ROC Area	n/a	**0.99**	0.95	**0.99**
Cars	Acc. (%)	94.93	98.17	n/a	**98.28**
	ROC Area	n/a	**1.00**	n/a	**1.00**
Faces	Acc. (%)	59.83	88.70	86.78	**89.57**
	ROC Area	n/a	**0.95**	0.91	**0.95**
Motorbikes	Acc. (%)	76.80	**88.99**	84.67	87.81
	ROC Area	n/a	**0.95**	0.92	0.94

Table 2: Results of binary classification between each of the four classes of Caltech-4 and the background clutter class. BoW: bag of word approach [22]. SVM: traditional SVM using global statistics. SVM-FS [15] is the SVM method that require strongly labeled data during training. Results of SVM-FS for the Cars class is displayed as n/a because of the unavailability of ground truth annotation.

4.2.1 A synthetic example

The data in this evaluation consists of 800 artificially generated examples of binary time series (400 positive and 400 negative). Some examples are shown in Fig. 5. Each positive example contains three long segments of fixed length with value 1. We refer to these as the foreground segments. Note that the end of a foreground segment may meet the beginning of another one, thus creating a longer foreground segment (see e.g. the bottom left signal of Fig. 5). The locations of the foreground segments are randomly distributed. Each negative example contains fewer than three foreground segments. Both positive and negative data are artificially degraded to simulate measurement noise: with a certain probability, zero energy values are flipped to have value 1. The temporal length of each signal is 100 and the length of each foreground segment is 10. We split the data into separate training and testing sets, each containing 400 examples (200 positive, 200 negative).

We evaluated the ability of our algorithm to discover automatically the discriminative segments in these weakly-labeled examples. We trained our localization-classification SVM by learning k-segmentations for values of k ranging from 1 to 20. Note that the algorithm has no knowledge of the length or the type of the pattern distinguishing the two classes. Tab. 3 summarizes the performance of our approach. Traditional SVM, based on the statistics of the whole signals, yields

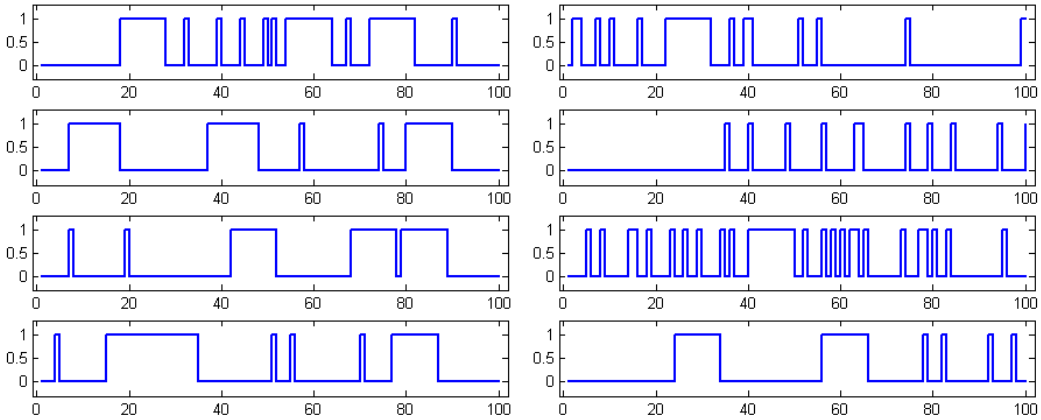

Figure 5: What distinguishes the time series on the left from the ones on the right? Left: positive examples, each containing three long segments with value 1 at random locations. Right: negative examples, each containing fewer than three long segments with value 1. All signals are perturbed with measurement noise corresponding to spikes with value 1 at random locations.

k	1	2	3 to 7	8	12	16	20
Acc.(%)	77.0	93.0	100	98.5	91.5	77.5	67.25
ROC Area	.843	.980	1.00	.998	.933	.793	.613

Table 3: Classification performance on temporal data using our approach. We show the accuracy rates and the ROC areas obtained using different values of k, the number of discriminative time intervals used by the algorithm. Here traditional SVM, based on the global statistics of the signals, yields an accuracy rate of 66.5% and an area under the ROC of 0.577.

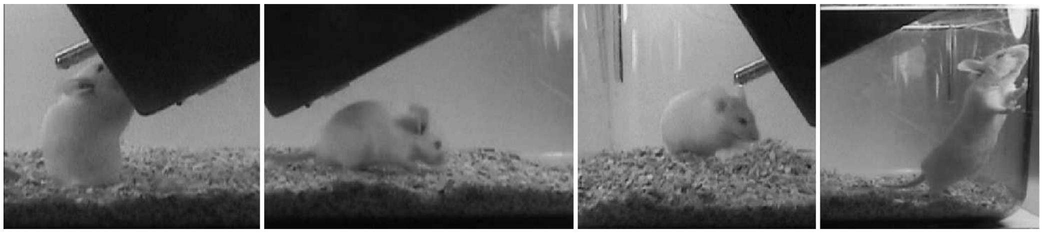

Figure 6: Example frames from the mouse videos.

an accuracy rate of 66.5% and an area under the ROC of 0.577. Thus, our approach provides much better accuracy than SVM. Note that the performance of our method is relatively insensitive to the choice of k, the number of discriminative time-intervals used for classification. It achieves 100% accuracy when the number of intervals are in the range 3 to 7; it works relatively well even for other settings. In practice, one can use cross validation to choose the appropriate number of segments. Furthermore, Tab. 3 reaffirms the need of using multiple intervals: our classifier built with only one interval achieves only an accuracy rate of 77%.

4.2.2 Mouse behavior

We now describe an experiment of mouse behavior recognition performed on a publicly available dataset[3]. This collection contains videos corresponding to five distinct mouse behaviors: drinking, eating, exploring, grooming, and sleeping. There are seven groups of videos, corresponding to seven distinct recording sessions. Because of the limited amount of data, performance is estimated using leave-one-group-out cross validation. This is the same evaluation methodology used by Dollár et al. [9]. Fig. 6 shows some representative frames of the clips. Please refer to [9] for further details about this dataset.

We represent each video clip as a set of *cuboids* [9] which are spatial-temporal local descriptors. From each video we extract cuboids at interest points computed using the cuboid detector [9]. To these descriptors we add cuboids computed at random locations in order to yield a total of 2500 points for each video (this augmentation of points is done to cancel out effects due to differing sequence lengths). A library of 50 cuboid prototypes is created by clustering cuboids sampled from

[3] http://vision.ucsd.edu/∼pdollar/research/research.html

Action	Dollár et al. [9]	1-NN	SVM	Ours
Drink	0.63	0.58	0.63	**0.67**
Eat	**0.92**	0.87	0.91	0.91
Explore	0.80	0.79	**0.85**	**0.85**
Groom	0.37	0.23	0.44	**0.54**
Sleep	0.88	0.95	**0.99**	**0.99**

Table 4: F1 scores: detection performance of several algorithms. Higher F1 scores indicate better performance.

training data using k-means. Subsequently, each cuboid is represented by the ID of the closest prototype and the frame number at which the cuboid was extracted. We trained our algorithm with values of k varying from 1 to 3. Here we report the performance obtained with the best setting for each class.

A performance comparison is shown in Tab. 4. The second column shows the result reported by Dollár et al. [9] using a 1-nearest neighbor classifier on histograms containing only words computed at spatial-temporal interest points. *1-NN* is the result obtained with the same method applied to histograms including also random points. *SVM* is the traditional SVM approach in which each video is represented by the histogram of words over the entire clip. The performance is measured using the F1 score which is defined as:

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}. \tag{10}$$

Here we use this measure of performance instead of the ROC metric because the latter is designed for binary classification rather than detection tasks [1]. Our method achieves the best $F1$ score on all but one action.

5 Conclusions and Future Work

This report proposes a novel framework for discriminative localization and classification from weakly labeled images or time series. We show that the joint learning of the discriminative regions and of the region-based classifiers leads to categorization accuracy superior to the performance obtained with supervised methods relying on costly human ground truth data. In future work we plan to investigate an unsupervised version of our approach for automatic discovery of object classes and actions from unlabeled collections of images and videos. Furthermore, we would like to extend our k-segmentation model to images in order to improve the recognition of objects having complex shapes.

Appendix – Proof of global optimality of Algorithm 1

Algo. 1 guarantees to produce a globally optimal solution for (8). Even stronger, the set $\mathcal{X}^m = \{I_1^m, \cdots, I_m^m\}$ produced by the algorithm is the set of best m intervals that maximize (8). This section sketches a proof by induction.

+) $m = 1$, this can be easily verified.
+) Suppose $\mathcal{X}^m$ is the set of best m intervals that maximize (8). We now prove that $\mathcal{X}^{m+1}$ is optimal for $m+1$ intervals. Assume the contrary, $\mathcal{X}^{m+1}$ is not optimal for $m+1$ intervals. There exist disjoint intervals $T_1, \cdots, T_{m+1}$ such that:

$$\sum_{i=1}^{m+1} f(T_i) > \sum_{i=1}^{m+1} f(I_i^{m+1}). \tag{11}$$

Because the way we construct $\mathcal{X}^{m+1}$ from $\mathcal{X}^m$, we have:

$$\sum_{i=1}^{m+1} f(I_i^{m+1}) = \sum_{i=1}^{m} f(I_i^m) + \max\{f(J_1), -f(J_2)\},$$

$$\text{where } J_1 = \arg\max_{J \in \mathcal{I}} f(J) \text{ s.t. } J \cap I_i^m = \phi \; \forall i, \tag{12}$$

$$J_2 = \arg\max_{J \in \mathcal{I}} -f(J) \text{ s.t. } J \subset I_i^m \text{ for an } i. \tag{13}$$

This, together with (11), leads to:

$$\max\{f(J_1), -f(J_2)\} < \sum_{i=1}^{m+1} f(T_i) - \sum_{i=1}^{m} f(I_i^m). \tag{14}$$

Consider the overlapping between $T_1, \cdots, T_{m+1}$ and $I_1^m, \cdots, I_m^m$, there are two cases.

• *Case 1:* $\exists j : T_j \cap I_i^m = \phi \ \forall i$. In this case, we have:

$$f(T_j) \le f(J_1) < \sum_{i=1}^{m+1} f(T_i) - \sum_{i=1}^{m} f(I_i^m), \tag{15}$$

$$\Rightarrow \sum_{i=1}^{m} f(I_i^m) < \sum_{i=\overline{1,m+1}, i \neq j} f(T_i). \tag{16}$$

This contradicts with the assumption that $\{\mathbf{I}_1^m, \cdots, \mathbf{I}_m^m\}$ is the set of best m intervals that maximize (8).

• *Case 2:* $\forall j, \exists i : T_j \cap I_i^m \neq \phi$. Since there are $m+1$ T_j's, and there are only m I_i^m's, there must exist one i s.t. I_i^m intersects with at least two of T_j's. Suppose l, l_1, l_2 are indexes s.t. $T_{l_1} \cap I_l^m \neq \phi$ and $T_{l_2} \cap I_l^m \neq \phi$. Furthermore, suppose T_{l_1}, T_{l_2} are consecutive intervals of T_j's (T_{l_1} precedes T_{l_2} and there is no T_j in between). Let $T_{l_1} = [t_{l_1}^-, t_{l_1}^+]$, $T_{l_2} = [t_{l_2}^-, t_{l_2}^+]$. Consider the interval $T = [t_{l_1}^+ + 1, t_{l_2}^- - 1]$. Because $T_{l_1} \cap I_l^m \neq \phi$ and $T_{l_2} \cap I_l^m \neq \phi$, T must be a subinterval of I_l^m, i.e. $T \subset I_l^m$. Hence

$$-f(T) \le -f(J_2) < \sum_{i=1}^{m+1} f(T_i) - \sum_{i=1}^{m} f(I_i^m), \tag{17}$$

$$\Rightarrow \sum_{i=1}^{m} f(I_i^m) < f(T) + \sum_{i=1}^{m+1} f(T_i), \tag{18}$$

$$\Rightarrow \sum_{i=1}^{m} f(I_i^m) < f(\underbrace{T_{l_1} \cup T \cup T_{l_2}}_{\text{an interval}}) + \sum_{i \neq l_1, l_2} f(T_i). \tag{19}$$

This contradicts with the assumption that $\{\mathbf{I}_1^m, \cdots, \mathbf{I}_m^m\}$ is the best set of m intervals that maximize (8).

Since both cases lead to a contradiction, $\mathcal{X}^{m+1}$ must be the best set of $m+1$ intervals that maximize (8). This completes the proof $\Box$.

Acknowledgments

This material is based upon work supported by the U.S. Naval Research Laboratory under Contract No. N00173-07-C-2040. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Naval Research Laboratory. Portions of this work were performed while Minh Hoai Nguyen and Lorenzo Torresani were at Microsoft Research Cambridge. The authors would like to thank Victor Lempitsky for useful discussion, Peter Gehler for pointing out related work, and Margara Tejera for helping with image annotation.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1475–1490, 2004.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Neural Information Processing Systems*, 2003.

[3] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *European Conference on Computer Vision*, 2008.

[4] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. *CVPR Workshop on Perceptual Organization in Computer Vision*, 2004.

[5] L. Cao and L. Fei-Fei. Spatial coherent latent topic model for concurrent object segmentation and classification. In *International Conference on Computer Vision*, 2007.

[6] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[7] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCV Workshop on Visual Surveillance & Performance Evaluation of Tracking and Surveillance*, 2005.

[10] C. Fanti, L. Zelnik-Manor, and P. Perona. Hybrid models for human motion recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscaled, deformable part model. In *Computer Vision and Pattern Recognition*, 2008.

[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition*, 2003.

[13] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object recognition and localization with stable segmentations. In *European Conference in Computer Vision*, 2008.

[14] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming (web page & software). http://stanford.edu/∼boyd/cvx, Oct. 2008.

[15] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: object localization by efficient subwindow search. In *Computer Vision and Pattern Recognition*, 2008.

[16] I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, 2003.

[17] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *British Machine Vision Conference*, 2003.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[19] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning*, 1998.

[20] M. H. Nguyen, J. Perez, and F. de la Torre. Facial feature detection with optimal pixel reduction SVMs. In *International Conference on Automatic Face and Gesture Recognition*, 2008.

[21] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, (3), 2008.

[22] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[23] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition*, 2006.

[24] H. Schweitzer. Utilizing scatter for pixel subspace selection. In *International Conference on Computer Vision*, 1999.

[25] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, volume 2, 2003.

[26] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[27] Z. Tu, X. Chen, A. Yuille, and S. Zhu. Image parsing: unifying segmentation, detection and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005.

[28] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[29] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[30] C. Yang and T. Lozano-Pérez. Image database retrieval with multiple-instance learning techniques. In *International Conference on Data Engineering*, 2000.

[31] S. X. Yu and J. Shi. Object-specific figure-ground segregation. In *Computer Vision and Pattern Recognition*, 2003.

[32] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.